Garcia, María del Carmen
Rapelli, Cecilia
Instituto de Investigaciones Teóricas y Aplicadas, de la Escuela de Estadística

EVALUACIÓN DEL EFECTO DE ESPECIFICAR INCORRECTAMENTE LA DISTRIBUCIÓN DE LOS EFECTOS ALEATORIOS EN UN MODELO NO LINEAL MIXTO.

1.- Introducción

Los modelos no lineales mixtos se usan, frecuentemente, para caracterizar longitudinalmente procesos biológicos. Para representar la variación poblacional, estos modelos expresan los parámetros específicos de las unidades en función de efectos fijos y aleatorios y para considerar la correlación entre las mediciones repetidas introducen errores intra unidad. Un supuesto usado habitualmente es el de distribución normal para los errores y los efectos aleatorios. El supuesto sobre estos últimos suele no ser acertado y su cumplimiento puede ser dificultoso de verificar con las herramientas estadísticas estándares. Debido a que la predicción de los efectos aleatorios depende tanto de los errores como de los efectos aleatorios, los gráficos usuales para comprobar el supuesto de normalidad no permiten diferenciar cual de los dos supuestos distribucionales es el incorrecto.

Varios autores han comprobado el efecto de la normalidad para los efectos fijos en los modelos no lineales mixtos, pero no se ha investigado el efecto que produce la falta de cumplimiento de este supuesto sobre la distribución de los efectos aleatorios predichos.

En este trabajo, a través de simulaciones, se investiga el impacto de la especificación incorrecta de la distribución sobre la estimación de los efectos fijos y la predicción de los efectos aleatorios y qué tan bien estos últimos recuperan la verdadera distribución subyacente.

2.- Modelo no lineal mixto

El modelo no lineal mixto para las observaciones de la unidad i, i=1,... N se puede expresar como,

$$\mathbf{Y}_i = f(\mathbf{X}_i, \boldsymbol{\beta}_i) + \mathbf{e}_i, \qquad (2.1)$$

donde, $\mathbf{Y}_i = [Y_{i1},...,Y_{in_i}]'$ es el vector (n_i x1) compuesto por las mediciones repetidas del i-

ésimo individuo, e Y_{ij} la observación realizada al i-ésimo individuo en el tiempo t_j, $j = 1,\ldots,n_i$,

$\boldsymbol{f}(\mathbf{X}_i,\boldsymbol{\beta}_i) = [f(\mathbf{x}_{i1},\boldsymbol{\beta}_i),\ldots,f(\mathbf{x}_{in_i},\boldsymbol{\beta}_i)]'$ siendo, f una función no lineal conocida que relaciona el vector de respuestas con el tiempo y otras posibles covariables intra-unidad ($\mathbf{X}_i$) y $\boldsymbol{\beta}_i$ es un vector específico del individuo que contiene los parámetros de la función no lineal.

El vector $\boldsymbol{\beta}_i$ se puede modelar, en una segunda etapa, como la suma de dos componentes, una fija o poblacional común a todos los sujetos y otra específica a cada sujeto,

$$\boldsymbol{\beta}_i = \mathbf{A}_i\boldsymbol{\beta} + \mathbf{B}_i\mathbf{b}_i . \qquad (2.2)$$

Los elementos del modelo no lineal mixto son, entonces,

$\mathbf{X}_i = \{\mathbf{x}_{ij}\}$: Matriz ($n_i$ x v) de diseño del i-ésimo individuo, $j = 1,\ldots,n_i$,

$\boldsymbol{\beta}_i$: Vector (rx1) de parámetros del sujeto i-ésimo,

$\boldsymbol{\beta}$: Vector (sx1) de efectos fijos,

$\mathbf{b}_i$: Vector (qx1) de efectos aleatorios,

$\mathbf{A}_i$: Matriz (rxs) de diseño para los efectos fijos,

$\mathbf{B}_i$: Matriz (rxq) de diseño para los efectos aleatorios.

Se supone que $\mathbf{b}_i$ y $\mathbf{e}_i$ son independientes con distribución,

$$\mathbf{b}_i \overset{\text{i.i.d.}}{\sim} N_q(\mathbf{0},\mathbf{D}) \qquad \mathbf{e}_i \overset{\text{i.i.d.}}{\sim} N_{n_i}(\mathbf{0},\boldsymbol{\Sigma}_i),$$

siendo, **D** la matriz de covariancias de los efectos aleatorios y $\boldsymbol{\Sigma}_i$, con la misma estructura para todos los individuos, la matriz de covariancias intra-individuos.

El modelo se puede estimar mediante el método de máxima verosimilitud. Condicional a los efectos aleatorios $\mathbf{Y}_i \sim N(f(.),\boldsymbol{\Sigma}_i)$, la verosimilitud para $\mathbf{Y}_i$ se puede obtener integrando una densidad normal con respecto a la distribución de los efectos aleatorios. Pero, maximizar la función de verosimilitud resultante es complicado por la presencia de una integral multidimensional en esta función. Para una estructura simple de los efectos aleatorios, la integración se puede realizar por cuadratura Gaussiana (Davidian y Gallant, 1993), o alguna otra técnica numérica sin demasiada dificultad. Existen varias alternativas para la estimación de la verosimilitud completa de los modelos no lineales mixtos (NLMM) que están basadas

en la expansión de Taylor de primer orden de la función f del modelo. La principal distinción entre esos métodos, denominados de linealización, reside en el punto alrededor del cual se hace la expansión. Ésta se puede realizar alrededor del valor esperado del vector de efectos aleatorios (0) (Sheiner y Beal, 1980), o alrededor de alguna estimación del vector de efectos aleatorios, usualmente llamado el mejor predictor lineal insesgado (EBLUP) (Lindstrom y Bates, 1990).

3.- Estudio de simulación

Para evaluar el efecto de especificar incorrectamente la distribución de los efectos aleatorios se lleva a cabo un estudio de simulación. Los datos se generan a partir de modelo que considera la función no lineal de Wood, cuyos tres parámetros representan el valor inicial de la respuesta, la tasa de ascenso hasta la máxima respuesta y de descenso desde el máximo. La elección de los parámetros para la simulación se inspira en los resultados del ajuste de un modelo para evaluar la evolución de la lactancia en vacas *Holando* (Garcia et. al., 2009). Los datos representan las mediciones de la producción de leche en 15 momentos a 120 vacas, registrados de acuerdo al número de parto (1º,2º,3 y más) al que corresponde esa lactancia. Se asume que los parámetros de la curva de Woods son una función lineal de tres efectos fijos y que los dos primeros poseen efectos aleatorios.

La expresión de la función ((2.1) y (2.2)) y los valores de los parámetros utilizados en la simulación son los siguientes,

$$Y_{ij} = \beta_{0i}\, t_{ij}^{\beta_{1i}}\, exp(-\beta_2\, t_{ij}) + e_{ij} \qquad (3.1)$$

$$\mathbf{e}_i \sim N\left(\mathbf{0}, \sigma^2 \mathbf{I}\right)$$

$$\begin{aligned} \beta_{0i} &= 24.0109 - 6.4001\, P_1 - 1.8808\, P_2 + b_{0i} \\ \beta_{1i} &= 0.3445 - 0.0967\, P_1 - 0.0549\, P_2 + b_{1i} \\ \beta_{2i} &= 0.1016 - 0.0347\, P_1 - 0.0112\, P_2 \end{aligned} \quad , \qquad (3.2)$$

siendo,

P_1=1 primer parto y 0 en otro caso.

P_2=1 segundo parto y 0 en otro caso.

β_{0i} el valor inicial de producción,

β_{1i} la tasa de ascenso

β_{2i} la tasa de descenso

$\mathbf{b}_i$: Vector (kx1) de efectos aleatorios

$\mathbf{e}_i$ Vector (n_ix1) de errores intra-grupo.

La variancia intra unidad es $\sigma^2 = 7.2540$.

El vector $\mathbf{b}_i = \begin{pmatrix} b_{0i} \\ b_{1i} \end{pmatrix}$ tiene distribución conocida con media cero y matriz de covariancias **D**

$\mathbf{D} = \begin{pmatrix} 33.2913 & -0.2965 \\ -0.2965 & 0.0111 \end{pmatrix}$.

Para generar los efectos aleatorios se consideran tres distribuciones,

1.- normal $\mathbf{b}_i \overset{iid}{\sim} N(\mathbf{0},\mathbf{D})$,

2.- t de Student $\mathbf{b}_i \overset{iid}{\sim} t_2(3,\mathbf{0},\mathbf{D})$ y

3.- Gamma multivariada con matriz de covariancias **D**.

La distribución normal representa el supuesto usual que se efectúa sobre los efectos aleatorios. La distribución t ejemplifica una situación en la que la distribución de los efectos aleatorios tiene colas más pesadas que la normal. La última representa un caso de distribución asimétrica.

Dos diferentes esquemas de generación se usan para investigar el comportamiento de los efectos fijos y los parámetros covariancia y para la predicción de los efectos aleatorios

Estimación efectos fijos y parámetros covariancia

Se generan tres conjuntos de datos de tamaño 120, suponiendo para cada uno una distribución diferente para los efectos aleatorios. Para cada conjunto se ajusta el modelo propuesto suponiendo normalidad de los efectos aleatorios y se registran las estimaciones de los parámetros y sus variancias Se repite el procedimiento 200 veces. Se calcula el promedio de las estimaciones de los parámetros y de sus variancias.

Predicción de efectos aleatorios

Se generan seis conjuntos de datos de 1200 unidades, 400 para cada parto, considerando distintas magnitudes para la variancia intra unidad (σ^2 = 30, 5 y 0.5). Cada conjunto corresponde a una combinación distribución efectos aleatorios y σ^2 . Para cada unidad se estiman los parámetros y se calculan los efectos aleatorios suponiendo normalidad. Se realizan histogramas y “box-plots” para visualizar la distribución de los mismos.

Se utilizó el procedimiento IML y la macro NLINMIX del sofyware estadístico SAS para ambos esquemas.

4.- Resultados

Los valores generados se utilizan para ajustar el modelo,

$$Y_{ij} = \beta_{0\,i}\, t_{ij}^{\beta_{1i}} \exp(-\beta_2\, t_{ij}) + e_{ij}$$

$$\begin{aligned}\beta_{0i} &= \beta_0 + \beta_{01}\ P_1 + \beta_{02}\ P_2 + b_{0i}\\ \beta_{1i} &= \beta_1 + \beta_{11}\ P_1 + \beta_{12}\ P_2 + b_{1i}\\ \beta_{2i} &= \beta_2 + \beta_{21}\ P_1 + \beta_{22}\ P_2\end{aligned}\quad ,$$

y calcular la predicción de los efectos aleatorios.

Se estima el modelo anterior usando el método de linealización alrededor de $\mathbf{b}_i$, que utiliza el supuesto de distribución normal para los efectos aleatorios. Se estiman los parámetros del mismo y se calcula la predicción de los efectos aleatorios, según corresponda.

Debido a que el procedimiento de estimación a veces puede no converger, los resultados que se presentan corresponden a la situación de convergencia del modelo; en caso contrario los resultados se omiten.

La tabla siguiente resume el comportamiento de los estimadores de los parámetros de efectos fijos, de covariancia y de la variancia del error en término de sesgo y los errores estándares de los estimadores.

Tabla 1 Resultados de la simulación para los parámetros de los efectos fijos y de covariancia

Parámetros	Valores reales	Distribución Normal			Distribución t, 3 gl			Distribución Gamma		
		Media	Sesgo	Std	Media	Sesgo	Std	Media	Sesgo	Std
β_0	24.01	23.92	-0.09	0.9389	23.91	-0.10	1.0232	23.94	-0.07	0.8322
β_{01}	-6.40	-6.55	-0.15	1.3543	-6.38	0.02	1.3941	-6.41	-0.01	1.3932
β_{02}	-1.88	-1.86	0.02	1.3137	-1.86	0.02	1.4119	-1.90	-0.02	1.2941
β_1	0.34	0.35	0.00	0.0246	0.35	0.01	0.0270	0.35	0.00	0.0242
β_{11}	-0.10	-0.09	0.00	0.0378	-0.09	0.00	0.0412	-0.10	0.00	0.0387
β_{12}	-0.05	-0.06	0.00	0.0387	-0.06	-0.01	0.0402	-0.06	0.00	0.0371
β_2	0.10	0.10	0.00	0.0039	0.10	0.00	0.0038	0.10	0.00	0.0037
β_{21}	-0.03	-0.03	0.00	0.0062	-0.03	0.00	0.0062	-0.03	0.00	0.0060
β_{22}	-0.01	-0.01	0.00	0.0058	-0.01	0.00	0.0057	-0.01	0.00	0.0055
D_{00}	33.29	33.24	-0.06	4.9125	28.57	-4.72	10.5017	32.31	-0.98	9.1906
D_{01}	-0.30	-0.31	-0.01	0.0747	-0.28	0.02	0.1813	-0.24	0.06	0.0638
D_{11}	0.01	0.01	0.00	0.0018	0.01	0.00	0.0149	0.01	0.00	0.0032
σ^2	7.25	7.20	-0.06	0.2524	7.21	-0.04	0.3103	7.19	-0.06	0.2525

La estimación de los efectos fijos es bastante robusta a desviaciones moderadas de la normalidad de la distribución de los efectos aleatorios. La razón de esto, como se discute en Hartford et al. (2000), se puede deber al hecho que se eliminan los resultados cuando el procedimiento no converge y quizás los conjuntos de datos no eliminados resultaron "más normales" que los omitidos. Los datos utilizados no proporcionan los mismos resultados para algunos de los parámetros de covariancia.

El gráfico 1 presenta la distribución muestral de los estimadores $\hat{\beta}_{0i} = B0, \hat{\beta}_{1i} = B1, \hat{\beta}_{2i} = B2$ obtenidos considerando diferentes distribuciones para los efectos aleatorios, mientras que el gráfico 2 muestra los "box plots" de los estimadores de covariancia y $\hat{\sigma}^2 = S2$.

Gráfico 1 "Box plots" de los estimadores de los efectos fijos de parámetros

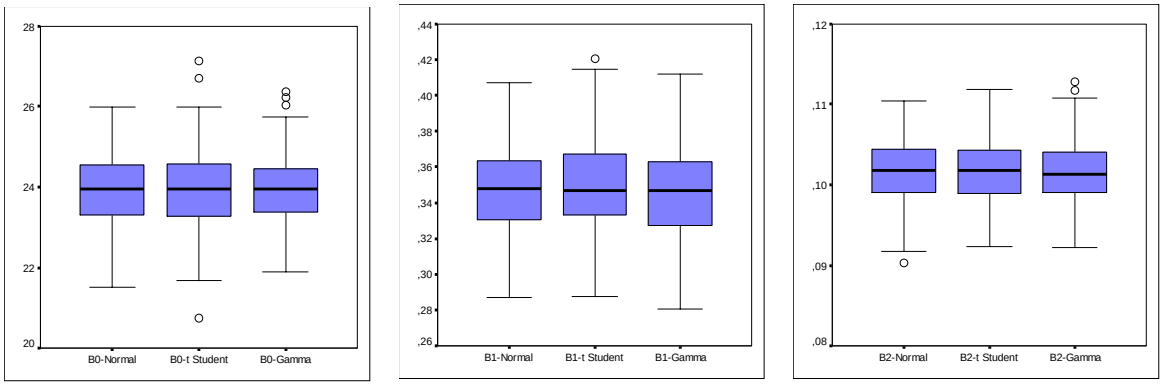
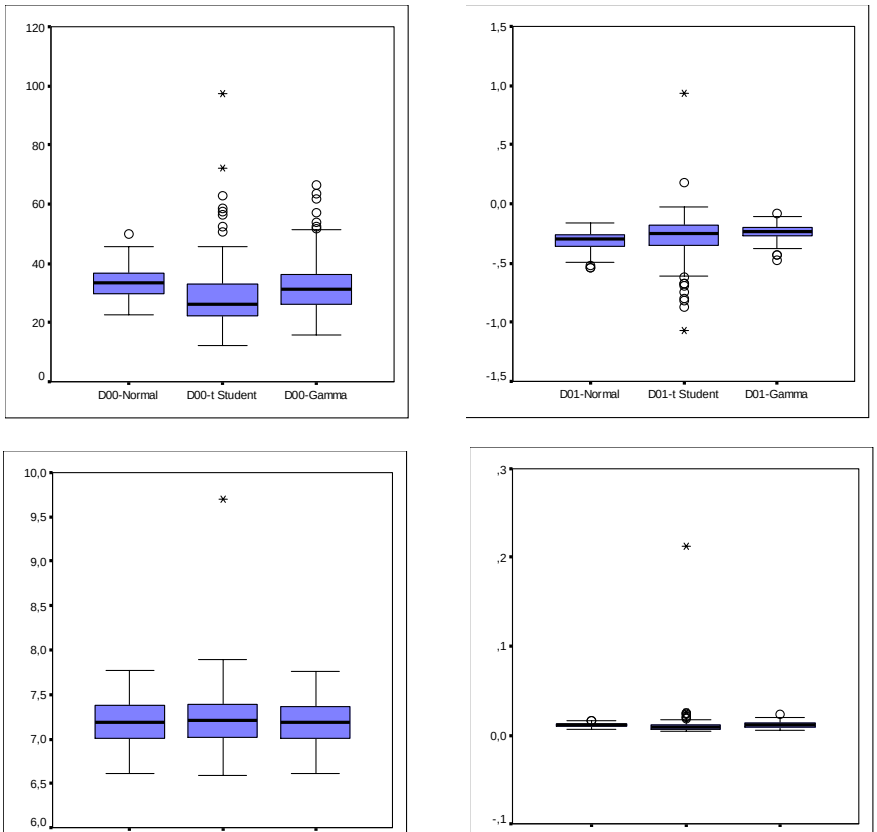


Gráfico 2 "Box plots" de los estimadores de los parámetros de covariancia



La ubicación de la mediana en el grafico 1 muestra la simetría de la mayoría de las distribuciones, excepto las correspondientes a $\hat{\boldsymbol{\beta}}_1$ para t-Student y Gamma que son algo asimétricas. Se observa, además, que las distribuciones no están demasiado dispersas y la presencia de valores atípicos para las no Gaussianas.

En el grafico 2 esta representación permite visualizar que la distribución muestral de los estimadores de **D** es bastante asimétrica en algunos de los casos en que los efectos aleatorios no son normales. Se observa, además, la presencia de valores atípicos.

Una conclusión general obtenida de los gráficos anteriores es que la estimación de la media de los parámetros está poco afectada por las desviaciones de la normalidad de los efectos aleatorios, mientras que la variabilidad de estos últimos está más afectada. Esta conclusión no es sorprendente debido a que se conoce que la media es relativamente robusta mientras que la variancia a menudo no lo es.

Los dos gráficos que se presentan a continuación permiten visualizar cómo se comporta la predicción de los efectos aleatorios cuando se asume un modelo gaussiano, cuando la verdadera distribución de los mismos no lo es.

El grafico 3 corresponde a la distribución muestral de la predicción de los efectos aleatorios asociado a $\hat{\boldsymbol{\beta}}_1$ asumiendo distribución gaussiana, siendo la verdadera distribución t- Student (izquierda) y gamma (derecha). El gráfico para el otro efecto aleatorio se presenta en el anexo.

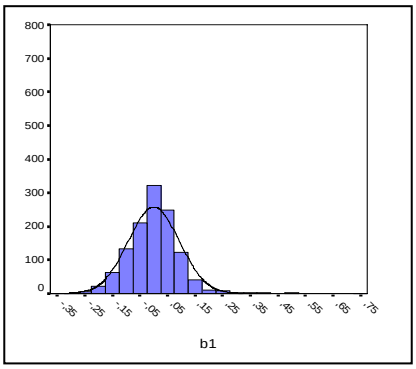
El histograma de la predicción de los efectos aleatorios muestra, para grandes valores de σ^2, que el supuesto de normalidad fuerza a los valores $\hat{b}_i$ a satisfacer ese supuesto.

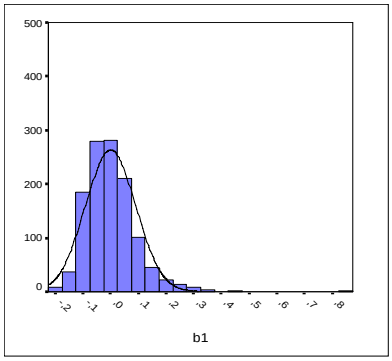
Gráfico 3. Distribución de la predicción del efecto aleatorio de la tasa de ascenso, para diferentes variancias

Distribucion t-Student **Distribucion Gamma**

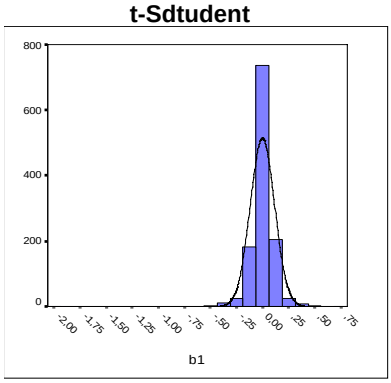
Verdadero

$\sigma^2 = 30$

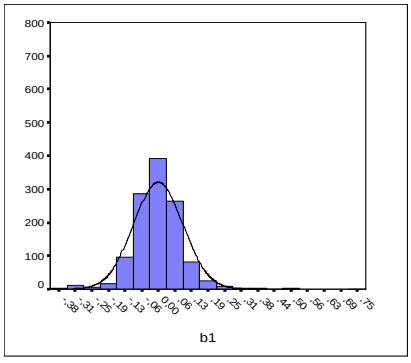


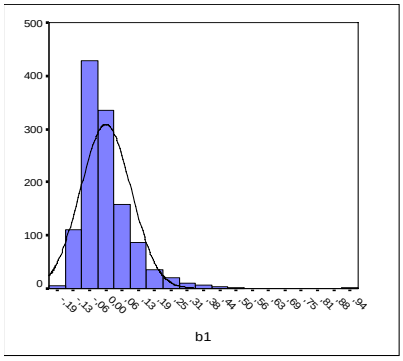


t-Sdtudent

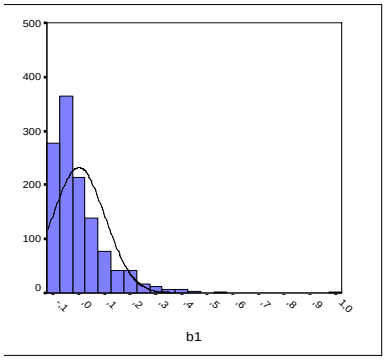


$\sigma^2 = 5$

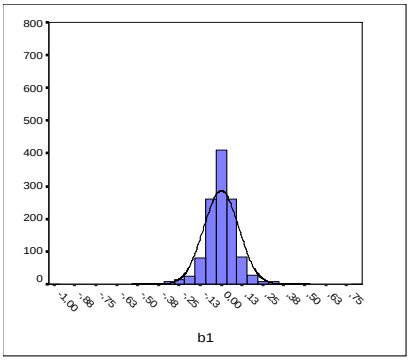




Gamma



$\sigma^2 = 0.5$



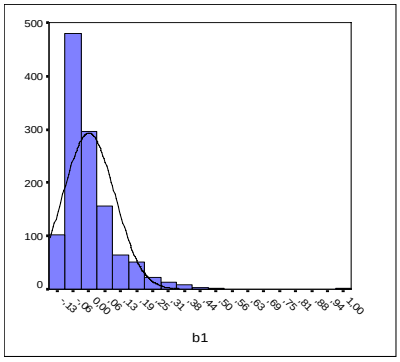
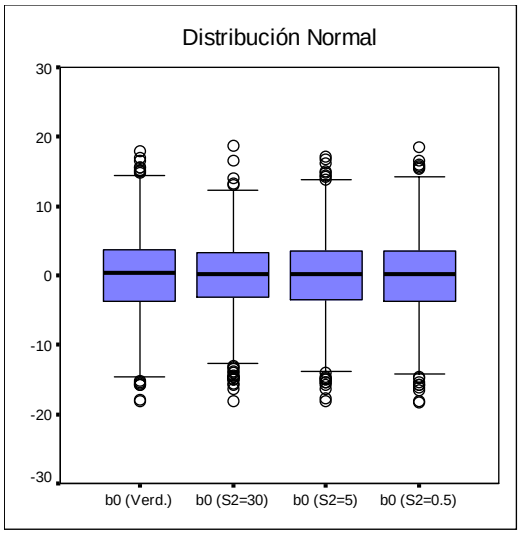
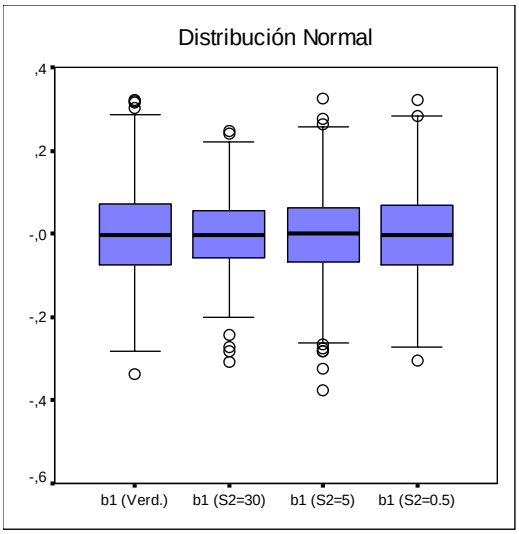
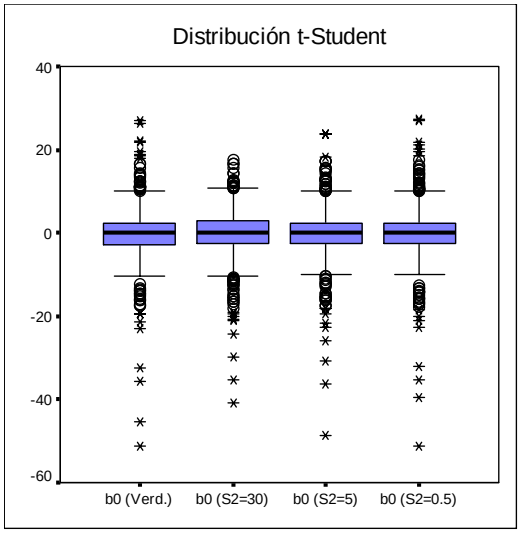
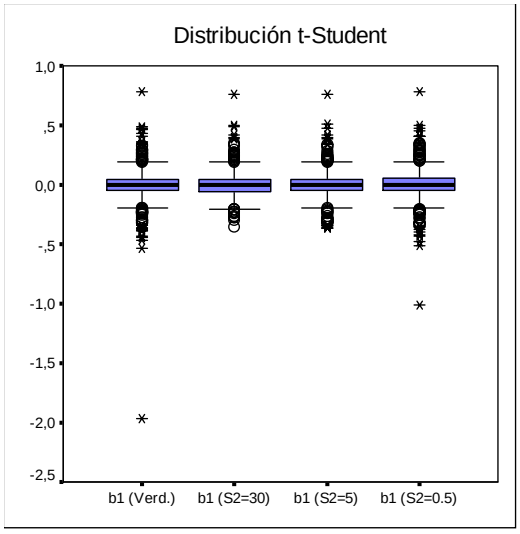


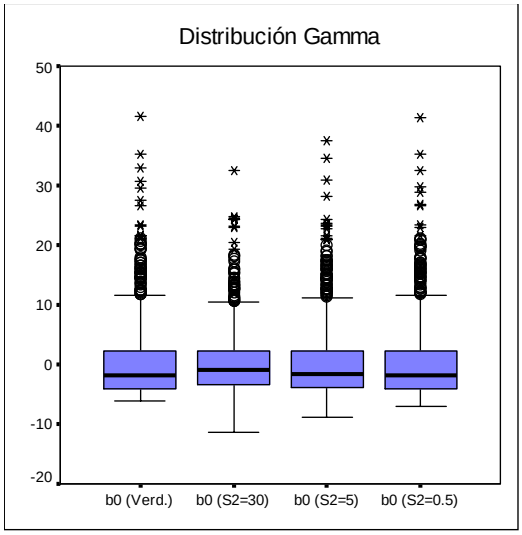
Gráfico 4. “Box plots” de los efectos aleatorios simulados para diferentes variancias

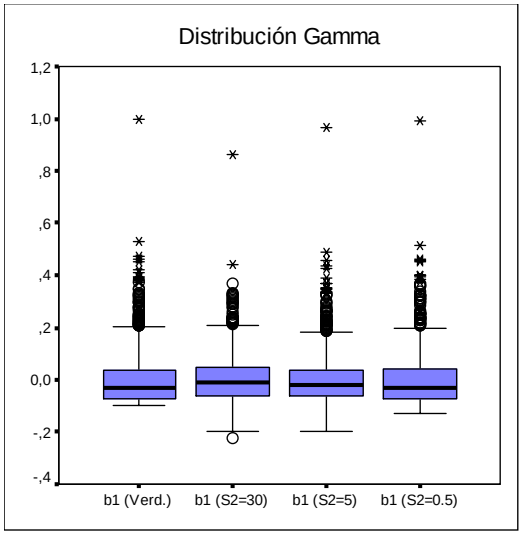












En el grafico 4 se observa que a medida que la variancia de error aumenta la distribución de la predicción de los efectos aleatorios no refleja la distribución de la población de la cual provienen.

5.- Consideraciones finales

En este trabajo se considera el problema que puede surgir cuando la distribución de los efectos aleatorios no es Gaussiana, la distribución supuesta por el método utilizado para estimar los parámetros de los modelos no lineales mixtos. El cumplimiento de este supuesto puede ser dificultoso de verificar con las herramientas estadísticas estándares, debido a que la predicción de los efectos aleatorios depende tanto de los efectos aleatorios como de los errores aleatorios y los gráficos probabilísticos normales pueden indicar sólo que las predicciones estandarizadas no tienen la distribución que se espera bajo el modelo asumido, pero no permiten diferenciar cual de los dos supuestos distribucionales es incorrecto.

A través de simulaciones, se investiga el impacto de la especificación incorrecta de la distribución sobre la estimación de los efectos fijos y la predicción de los efectos aleatorios y qué tan bien estos últimos recuperan la verdadera distribución subyacente.

Se destaca que:

- La estimación de los efectos fijos no está comprometida por la falta de normalidad, sin embargo se observa sesgo en la estimación de algunos de los parámetros de covariancia.
- Aunque las predicciones pueden ser sensibles a la distribución asumida, en general en este estudio se observa que las predicciones están algo afectadas por el alejamiento de este supuesto.
- La variancia del error parece jugar un papel importante en la forma de la distribución de la predicción de los efectos aleatorios.

En resumen, la estimación de los parámetros del modelo marginal en este modelo no lineal mixto particular está levemente influenciado por el supuesto de normalidad de los efectos aleatorios. Pero esto no se cumple totalmente para la predicción de los efectos aleatorios.

Los resultados obtenidos permiten comprender mejor las consecuencias de incumplir los supuestos del procedimiento de estimación, pero no se pueden extraer conclusiones generales. Resulta necesario realizar posteriores investigaciones que pueden estar basadas en este trabajo preliminar.

REFERENCIAS BIBLIOGRÁFICAS

Davidian, M., Gallant, A. 1994. Nlmix, a program for maximum likelihood estimation for the nonlinear mixed effects model with a smooth random effects density. North Carolina State University,Raleigh.

Davidian, M., Giltinan, D., 1993. Some general estimation methods for nonlinear mixed effects models. J*. Biopharm. Stat*. 3, 23-55.

Davidian, M., Giltinan, D., 1995. Nonlinear models for repeated measurement data. Chapman & Hall, New York.

Garcia, M. del C., Rapelli, C., Cuatrín, A. 2010 Multilevel nonlinear mixed model for modeling and choosing a lactaction curve BIOCELL vol. 34 ISSN 0327-9545 (abstract)

Hartford, A. y Davidian, M. 2000 Consequences of misspecifying assumptions in nonlinear mixed effects models. Computational Statistics & Data Analysis 34, 139-164

Lindstrom, M., Bates, D., 1990. Nonlinear Mixed effects models for repeated measures data. Biometrics 46, 673-687.

Littell, R., Milliken, G., Stroup, W., Wolfinger, R. 1996. SAS System for mixed models. SAS Institute Inc., Cary, NC.

SAS/IML software: usage and reference. version 8.(SAS Institute Inc., Cary, NC.)

Sheiner, L., Beal, S., 1980. Evaluation of methods for estimating population pharmacokinetic parameters. I. Michaelis-Menten model: routine clinical pharmacokinetic data. J. Pharmacokin. Biopharm. 8, 553-571.

Verbecke, G. and Lesaffre, E. 1996. A linear mixed-effects model with heterogeneity in the random-effects population. Journal of the American Statistical Association 91, 217–221.

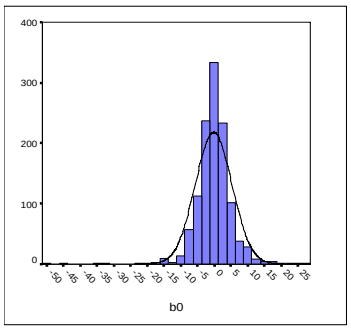
Verbecke, G. and Lesaffre, E. 1997. The effect of misspecifying the random-effects distribution in linear mixed models for longitudinal data. Computational Statistics and Data Analysis 23, 541–556.

Vonesh, E., Chinchilli, V. 1997 Linear and nonlinear models for the repeated measurements. Marcel Dekker.

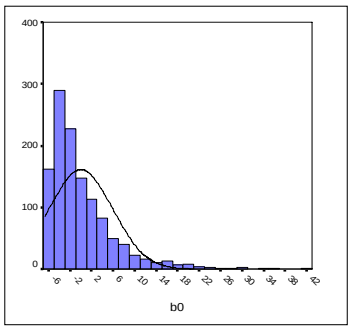
ANEXO

Gráfico A1: Distribución de la predicción del efecto aleatorio del valor inicial, para diferentes variancias

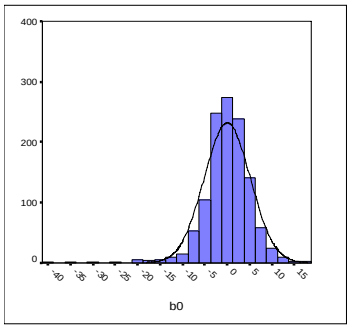
t-Student
Verdaderos



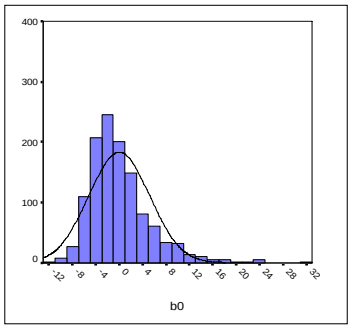
Gamma
Verdaderos



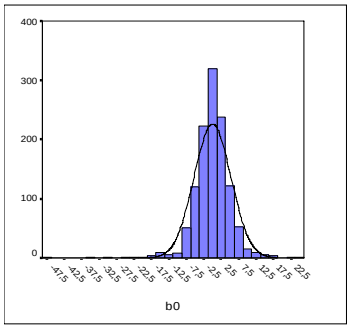
σ^2=30



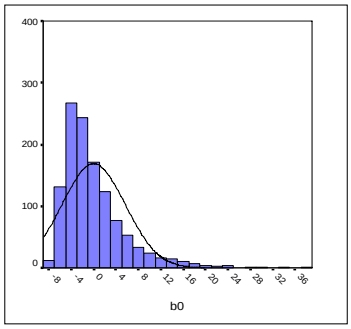
σ^2=30



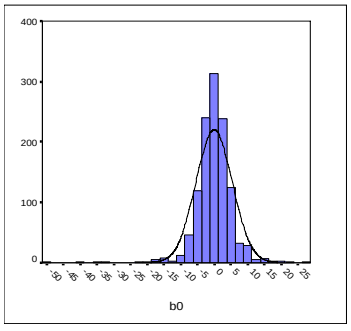
σ^2=5



σ^2=5



σ^2=0.5



σ^2=0.5